Biometrics, 2024, 80(2), ujae019
https://doi.org/10.1093/biomtc/ujae019
Biometric Practice

OXFORD

Case weighted power priors for hybrid control analyses with time-to-event data

Evan Kwiatkowski [1,*], Jiawen Zhu[2], Xiao Li[2], Herbert Pang[2], Grazyna Lieberman[2], Matthew A. Psioda [3]

[1]Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston, 1200 Pressler St, Houston, TX 77030, USA, [2]Department of Biostatistics, Genentech, South San Francisco, CA 94080, USA, [3]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

[*]Corresponding author: Evan Kwiatkowski, Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston,1200 Pressler St, Houston, Texas 77030, USA (evan.k.kwiatkowski@uth.tmc.edu)

ABSTRACT

We develop a method for hybrid analyses that uses external controls to augment internal control arms in randomized controlled trials (RCTs) where the degree of borrowing is determined based on similarity between RCT and external control patients to account for systematic differences (e.g., unmeasured confounders). The method represents a novel extension of the power prior where discounting weights are computed separately for each external control based on compatibility with the randomized control data. The discounting weights are determined using the predictive distribution for the external controls derived via the posterior distribution for time-to-event parameters estimated from the RCT. This method is applied using a proportional hazards regression model with piecewise constant baseline hazard. A simulation study and a real-data example are presented based on a completed trial in non-small cell lung cancer. It is shown that the case weighted power prior provides robust inference under various forms of incompatibility between the external controls and RCT population.

KEYWORDS: Box's *p*-value; historical control; power prior; prior-data conflict; real-world data.

1 INTRODUCTION

Randomized controlled trials (RCTs) are the gold standard for generating evidence in the confirmatory trial setting. If there is existing data on the standard of care used in the control group, then a hybrid control design that makes use of this data can generate more evidence than a RCT alone, or a similar amount of evidence in conjunction with a RCT with fewer subjects. External data on the control group could be available from real-world data (RWD), which could introduce confounding due to selection bias and may have a higher incidence of measurement error compared to historical trials due to variability in data collection procedures outside of the trial setting. Therefore, it is necessary for a hybrid control design to balance the risks and rewards in using external information, particularly with regard to possible systematic differences between data sources.

RWD are often used to construct an informative prior distribution for parameters common with the RCT data model, such as prognostic covariates separate from the treatment effect. An issue with using an informative prior is the possibility of incompatibility between the prior and the observed data, referred to as prior-data conflict. The prior predictive distribution informs which observable data are plausible based on the prior distribution and can be used to assess the compatibility of a prior and the data. When RWD are surprising (i.e., unlikely in a probabilistic sense) based on a prior predictive distribution derived from the RCT data, this signals that something may be systematically different between the generative process for the RWD and the RCT data (Lek and Van De Schoot, 2019). Box (1980) describes how the prior predictive distribution can be used to assess the compatibility of a prior and the data (i.e., Box's *p*-value). This can be used to identify priors that conflict with the observed data (Evans and Moshonov, 2006), a concept that has been used recently in adaptive trial design (e.g., Psioda and Xue (2020); Kwiatkowski et al. (2022)).

Bayesian dynamic borrowing approaches permit the degree of borrowing to depend on the heterogeneity between the current and external data (Viele et al. 2014). The power prior (Ibrahim and Chen, 2000) assumes that the two sources of data are exchangeable, and that differences between observed external and randomized control response rates are solely attributable to known covariate effects and sampling variation (also known as "conditional exchangeability of controls" (Psioda and Ibrahim, 2018)). The commensurate prior (Hobbs et al. 2011) assumes that differences between response rates are attributable to different parameters for the covariate effects in each data source. The meta-analytic predictive (MAP) approach (Neuenschwander et al. 2010) assumes that differences in response rates are attributable to between-trial heterogeneity and is better suited for considering multiple sources of external data (Dejardin et al. 2018). These existing approaches provide discounting based on

Received: July 6, 2022; Revised: December 31, 2023; Accepted: February 27, 2024

a collective assessment of the entire external dataset. Considering the information bias (e.g., measurement error) that may be present in RWD, the ability to individually discount specific observations may be a more appropriate approach for incorporating external evidence.

Propensity score methods have been used in conjunction with Bayesian dynamic borrowing to account for imbalance in baseline covariate distributions (Wang et al. 2019; 2022; Chen et al. 2022), although the same benefits are often present with conventional covariate adjustment (Fu et al. 2023). Therefore, an analysis method that permits discounting individual observations should use individual-level covariates in assessing compatibility.

In this paper, we develop a method for case-specific weighting of external controls. We assess the compatibility of each external control individually based on assessments of prior-data conflict using Box's p-value. This case-specific weighting can capture varying amounts of incompatibility among the external controls. We aim to have this method maintain traditional type I error control and to perform well in scenarios where the RWD is biased due to confounding which effects all the external controls equally. We use the commensurate prior for comparison since it explicitly includes a commensurability parameter that can represent drift in baseline hazard in the setting of survival analysis.

The motivating example for the simulation studies and real data analysis is a completed trial in non-small cell lung cancer (Rittmeyer et al. 2017) for which there are relevant potential external controls, which is reimagined as if a hybrid control arm were included as a part of the trial's design. The reimagined design uses external controls from RWD, which may have the confounding due to selection bias and measurement error for which individual-level discounting would be appropriate. It is instructive to consider a well-powered RCT with a large amount of external controls available for hybrid analyses since this enables simulation studies using random subsets of trial and external subjects to be compared to the known result from the RCT. The possible improvements in trial efficiency can then be measured through the operating characteristics of a design with fewer randomized subjects than were used in the actual trial, wherein we demonstrate the advantages of our method over existing borrowing approaches (e.g., higher power, lower mean squared error).

The rest of this paper is organized as follows: in Section 2, we define a compatibility function for the external controls that is used to determine the case-specific weights and construct the case weighted power prior. In Section 3, we provide a simulation study demonstrating the case weighted power prior under different types of confounding in the external data. In Section 4, we provide an analysis using the case weighted power prior on a real data set. We close the paper with some discussion in Section 5.

2 METHODS

2.1 External likelihood

For RCT subject i, y_{1i} is the observation time, ν_{1i} is the event indicator, $\mathbf{x}_{1i}$ is the covariate vector, and z_{1i} is the binary treatment indicator. The observation time y_{1i} is given as $y_{1i} = \min\{t_{1i}, c_{1i}\}$, where t_{1i} is the event time and c_{1i} is the censoring time. Denote the RCT data by $\mathbf{D}_1 = \{(y_{1i}, \nu_{1i}, \mathbf{x}_{1i}, z_{1i})\colon i = 1,...n_1\}$, where n_1 is the number of RCT subjects.

We consider a proportional hazards model with baseline hazard parameters $\boldsymbol{\lambda}$, covariate effect regression parameters $\boldsymbol{\beta}$, and treatment effect γ, with all unknown parameters denoted by $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \boldsymbol{\beta}, \gamma\}$. Let $\mathbf{x}_i$ and z_i denote the covariate vector and binary treatment indicator for RCT subject i, respectively. The hazard for RCT subject i is represented as $h_i(t|\boldsymbol{\theta}) = h_0(t|\boldsymbol{\lambda})\exp(\mathbf{x}_i'\boldsymbol{\beta} + z_i\gamma)$, and the hazard for external control j is represented as $h_j(t|\boldsymbol{\lambda}, \boldsymbol{\beta}) = h_0(t|\boldsymbol{\lambda})\exp(\mathbf{x}_j'\boldsymbol{\beta})$. The same proportional hazards model (i.e., baseline hazard and covariate effect) for the outcome is used for the RCT and external control data to allow compatibility assessments to be made based on these shared parameters. Partition the time axis into K intervals using $0 = \tau_0 < \tau_1 < \tau_2 <... < \tau_K = \infty$, and let $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_K)$. The partition of the time axis is chosen to capture important changes in the hazard rate while not introducing too many parameters. While there are formal regularization approaches for determining the partition (e.g., Bouaziz and Nuel (2016)), henceforth we suggest the partition induced by a prespecified number of quantiles for the events, which is the most commonly used method in practice. Let $\mathcal{I}_k = (\tau_{k-1}, \tau_k]$. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)^T$. The baseline hazard $h_0(t|\boldsymbol{\lambda})$ is taken as piecewise constant with $h_0(t|\boldsymbol{\lambda}) = \lambda_k$ for $t \in \mathcal{I}_k$. This form of the baseline hazard is chosen to provide additional flexibility over parametric models, such as the exponential or Weibull model, which makes it more attractive for practical use. There is also justification for the piecewise constant baseline hazard based on the theoretical connection to the Cox partial likelihood when each interval contains a single event (Ibrahim et al. 2001).

Denote the external control data by $\mathbf{D}_0 = \{(y_{0j}, \nu_{0j}, \mathbf{x}_{0j})\colon j = 1,..., n_0\}$, with all quantities analogous to the RCT data. Define the weighted likelihood for the external controls with both subject- and interval-specific weights by

$$\prod_{j=1}^{n_0} \mathcal{L}(\boldsymbol{\beta}, \lambda|\mathbf{D}_{0j}, \mathbf{a}_j) = \prod_{j=1}^{n_0}\left\{(\lambda_{K_j}\exp(x_j^T\boldsymbol{\beta}))^{a_{j,K_j}\nu_j} \times \prod_{k=1}^{K_j}\exp\left\{-a_{j,k}\lambda_k H_{j,k}\exp(x_j^T\boldsymbol{\beta})\right\}\right\}, \quad (1)$$

where $\mathbf{D}_{0j} = \{(y_{0j}, \nu_{0j}, \mathbf{x}_{0j})\}$ is the data for external control j, $\mathbf{a}_j = \{a_{j,1}, \ldots, a_{j,K_j}\}$ is a vector of interval-specific weights for external control j, $K_j \in \{1,..., K\}$ is the index for the interval such that $y_j \in \mathcal{I}_{K_j}$, and $H_{j,k}$ represents at-risk time during interval $\mathcal{I}_j$ for external control j for $k = 1,..., K_j$. Throughout we assume that, conditional on covariates, the censorship times are independent of the event times for both the RCT and external data.

2.2 Case weights

To determine the value of weights $\mathbf{a}_j = \{a_{j,1}, \ldots, a_{j,K_j}\}$ for external control j, we assess the compatibility of the time at risk in interval k, $H_{j,k}$, relative to its predictive distribution derived from the RCT data to determine whether the value is extreme relative

to what would be expected. The computation of the weights $\mathbf{a}_j$ is related to the memoryless property of the exponential distribution, which states that the probability a subject experiences an event after time t (given they do not have an event before that point) does not depend on the probability that they experience an event prior to time t. This property also applies to the piecewise exponential distribution and can be used for straightforward simulation of event times by simulating data for each time interval between cutpoints in $\boldsymbol{\tau}$ from independent exponential distributions. This affords us the ability to create a separate case weight for each interval for each external control by assessing the compatibility of the time at risk in the interval relative to its predictive distribution.

To compute this predictive distribution, it is necessary to specify and estimate a model for random censoring as it occurs in the RWD because the time at risk in an interval is a function of both the event and censorship distribution. This model for censoring does not need to be specified for the RCT data (assuming event and censoring times are independent), since the compatibility of external control observation times will be assessed with respect to parameters in the distribution for event times. The hazard for censorship for external control j may be represented as $h_j^c(c|\boldsymbol{\lambda}^c, \boldsymbol{\beta}^c) = h_0(c|\boldsymbol{\lambda}^c)\exp(\mathbf{x}_j'\boldsymbol{\beta}^c)$, where $\boldsymbol{\lambda}^c = \{\lambda_1^c, \ldots, \lambda_K^c\}$ are the baseline hazard parameters, and $\boldsymbol{\beta}^c$ are covariate effect regression parameters. As was done with the event-driven partition for the piecewise hazard in the time-to-event model, the time axis for instances of censoring is induced by a pre-specified number of quantiles. Therefore, the baseline hazard for occurrences of censoring is flexible to reflect censoring patterns likely to be seen in practice, such as infrequent censoring at the start of follow-up and substantial censoring after a certain amount of time.

Let y_j^{rep} be defined as replicated data (Gelman et al. 2013) that could have been observed using the same model and value for θ that produced the randomized control data, the same censoring model that produced stochastic censoring in the external controls, and the same covariate vector $\mathbf{x}_j$ as external control j. The predictive distribution for y_j^{rep} is given by

$$p(y_j^{\text{rep}}|\mathbf{D}_1, \mathbf{D}_0) = \int p(y_j^{\text{rep}}|\mathbf{x}_j, \boldsymbol{\lambda}, \boldsymbol{\lambda}^c, \boldsymbol{\beta})\pi(\boldsymbol{\lambda}, \boldsymbol{\beta}|\mathbf{D}_1) \times \pi(\boldsymbol{\lambda}^c|\mathbf{D_0})d\boldsymbol{\lambda}d\boldsymbol{\lambda}^c d\boldsymbol{\beta}, \quad (2)$$

where $p(y_j^{\text{rep}}|\boldsymbol{x}_j, \boldsymbol{\lambda}, \boldsymbol{\lambda}^c, \boldsymbol{\beta})$ is the density of the observation time for the jth external control, $\pi(\boldsymbol{\lambda}, \boldsymbol{\beta}|\mathbf{D}_1)$ is the posterior distribution for $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ based on the randomized control data, and $\pi(\boldsymbol{\lambda}^c|\mathbf{D_0})$ is the posterior distribution for $\boldsymbol{\lambda}^c$ based on the RWD. For a particular interval $\mathcal{I}_k$ (assuming external control j is at risk in interval k), the predictive distribution from equation (2) becomes

$$p(y_{j,k}^{\text{rep}}|\mathbf{D}_1, \mathbf{D}_0) = \int p(y_{j,k}^{\text{rep}}|\mathbf{x}_j, \lambda_k, \lambda_k^c, \boldsymbol{\beta})\pi(\boldsymbol{\lambda}, \boldsymbol{\beta}|\mathbf{D}_1) \times \pi(\boldsymbol{\lambda}^c|\mathbf{D_0})d\lambda_k d\lambda_k^c d\boldsymbol{\beta}. \quad (3)$$

It is our objective to use the value of the predictive density from equation (3) to assess compatibility of the observed RWD, such that observation times that are extreme relative to their predictive distribution will have comparatively lower predictive density values. Using a proportional hazards model with piecewise constant baseline hazard, it is necessary to transform the predictive distribution from equation (3) so that the mode does not occur at time zero. Using the predictive density from equation (3) would only allow observations that are higher than anticipated to be determined as incompatible based on their predictive density value. This transformation will use a function t such that $w_{j,k}^{\text{rep}} = t(y_{j,k}^{\text{rep}}|\mathbf{D}_1, \mathbf{D}_0)$ is approximately normally distributed. The function $t(x) = \log(x)$ is used so that the mode of the transformed density occurs near the expected value of the observation time, allowing for observation times that are either lower or higher than anticipated to be evaluated as more extreme. The weight $a_{j,k}$ is assigned as the probability of observing data as or more extreme (i.e., less likely) than the observed external control value $w_{j,k} = t(y_{j,k})$, and is an implementation of Box's p-value (Box, 1980). Formally, this is given by

$$a_{j,k} = \Pr\left[p(w_{j,k}^{\text{rep}}) \leq p(w_{j,k})\right], \quad (4)$$

where the probability (i.e., expectation) is taken with respect to the density $p(w_{j,k}^{\text{rep}})$. If there is perfect compatibility of the RCT controls and external controls, then the weights $a_{j,k}$ will be uniformly distributed since the shared parameters are equivalent and the posterior predictive distribution is continuous (Gelman et al. 2013). Further details on the computational implementation are in Web Appendix A.

2.3 Case weighted power priors

We use the weights $a_{j,k}$ defined using equation (4) to create a case weighted power prior as a generalization of the fixed-weight power prior $\pi_0(\boldsymbol{\theta}|\mathbf{D}_0, a_0) \propto [\mathcal{L}(\boldsymbol{\theta}|\mathbf{D}_0)]^{a_0}\pi_0(\boldsymbol{\theta})$. We replace $[\mathcal{L}(\boldsymbol{\theta}|\mathbf{D}_0)]^{a_0}$ in the fixed-weight power prior with the weighted likelihood for the external controls with both subject- and interval-specific weights from equation (1) with the addition of a calibration function which influences the operating characteristics of the analysis. This calibration function $h(a_{j,k}, \overline{A})$ is applied to each weight $a_{j,k}$ individually, and also is based on the average case weight for all external controls $\overline{A} = \sum_{j=1}^{n_0}\sum_{k=1}^{K_j} a_{j,k}/\sum_{j=1}^{n_0} K_j$. The average case weight $\overline{A}$ is used to detect dataset-level incompatibility among the external controls. Other functions of the weights could be used instead of the average (e.g., quantiles), which would then be compared to their expected values under perfect compatibility for calibration. The calibrated weighted likelihood becomes

$$\prod_{j=1}^{n_0}\mathcal{L}(\boldsymbol{\beta}, \lambda|\mathbf{D}_{0j}, h(\mathbf{a}_j, \overline{A})) = \prod_{j=1}^{n_0}\left\{(\lambda_{K_j}\exp(x_j^T\boldsymbol{\beta}))^{h(a_{j,K_j},\overline{A})v_j} \times \prod_{k=1}^{K_j}\exp\left\{-h(a_{j,k}, \overline{A})\lambda_k H_{j,k}\exp(x_j^T\boldsymbol{\beta})\right\}\right\}. \quad (5)$$

The resulting power prior using the likelihood in equation (5) with $h(a_{j,k}, \overline{A}) = f_p(a_{j,k})$ defines the case weighted power prior, where the function f_p is referred to as the case weight

shrinkage function. The parameter p controls the degree to which the case weights are tempered toward the constant 0.5 (their expected value under perfect compatibility) to counterbalance the modest type I error rate inflation which would arise from using $a_{j,k}$ directly in equation (5) (i.e., using untransformed case weights) with the conservative type I error rate of the fixed weight power prior, in order to produce an analysis with controlled type I error rate at the nominal level. See Web Appendix B and C for additional explanation.

The resulting power prior using the the likelihood in equation (5) with $h(a_{j,k}, \overline{A}) = f_p(a_{j,k})g_c(\overline{A})$ defines the discounted case weighted power prior, where the function g_c is referred to as the uniform discounting function. The function g_c is based on a predetermined level of maximum tolerated type I error rate (e.g., 0.15 is used henceforth) in the event that there is a shift in baseline hazard for all external controls. A similar process was implemented by Psioda et al. (2018), and also could be framed as a predetermined maximum level of power reduction for a shift in baseline hazard for all external controls. The parameter c controls the degree to which all case weights are reduced in value, which would tend toward a no-borrowing design based on the difference of $\overline{A}$ with 0.5 (its expected value under perfect compatibility).

The calibration procedure considers the given model (e.g., proportional hazards model with specific set of covariates) and given sample sizes for RWD and RCT data under the assumption of compatible external controls, and is therefore separate from the actual RWD outcomes (see details in Web Appendix D). The calibrated weights are also used to derive the case weighted commensurate prior which serves as a comparison method (see details in Web Appendix E).

3 SIMULATION STUDIES

3.1 Simulation setup

As a motivating example, we consider NCT02008227 (OAK study) Rittmeyer et al. (2017), a global, multicenter, open-label, randomized and controlled study, which evaluated the efficacy and safety of atezolizumab compared with docetaxel in participants with locally advanced or metastatic non-small cell lung cancer (NSCLC) after failure with platinum-containing chemotherapy. Among 850 participants randomized 1:1, an analysis of overall survival using Cox partial likelihood yields an estimated hazard ratio of 0.73 with 95% CI (0.62-0.86) in favor of atezolizumab.

We consider RWD from the nationwide (EHR)-derived longitudinal Flatiron Health database, comprised of de-identified patient-level structured and unstructured data curated via technology-enabled abstraction originating from ∼280 US cancer clinics (∼800 sites of care) (Ma et al. 2020; Birnbaum et al. 2020). Existing research has used Flatiron Health databases for external control analyses in oncology studies (Ventz et al. 2019; Lewis et al. 2019; Schmidli et al. 2019). We consider 526 external controls that meet OAK inclusion/exclusion criteria, henceforth referred to as NSCLC RWD.

Compatibility is assessed based on models which adjust for covariate effects; therefore, the methods considered take into account differences in measured characteristics. In fact, the data sources have different distributions of observed covariates (i.e., external controls average age 67.1 vs. RCT 63.2, external controls male 55.6% vs. RCT 66.2%), and the covariate effects are independently estimated from the fitted model parameters from the respective data source and therefore are distinct. We adjust for sex and age as measured covariates thought to be of prognostic value, and introduce confounding through a covariate (i.e., unobserved confounder), which represents systematic differences in the two hazards that are not explainable by measured covariates. We considered a hazard model for RCT subject i given by $h_i(t|\boldsymbol{\theta}) = h_0(t|\boldsymbol{\lambda})\exp(\text{age}_i\beta_1 + I(\text{sex} = \text{male})_i\beta_2 + z_i\gamma)$, and let the hazard for external control j be $h_j(t|\boldsymbol{\lambda}, \boldsymbol{\beta}) = h_0(t|\boldsymbol{\lambda})\exp(\text{age}_j\beta_1 + I(\text{sex} = \text{male})_j\beta_2 + x_{3j}\beta_3)$, where x_{3j} is the confounding covariate for external control j.

We consider three types of confounding based on the confounding covariate x_{3j}. "Partial contamination" occurs when the unobserved confounder affects a subset of the external controls, indicating nonexchangeability of a latent subpopulation, which could arise from a data quality issue (e.g., site-specific measurement error) in a clinical trial or incorrect information in an external control's electronic health record, which provides crucial data in a RWD cohort. This type of measurement error is likely common, but few methods are available to explicitly account for it. In particular, $x_3 = \log(2^m)$, with $\Pr(m = 0) = 0.68$ indicating no confounding and $\Pr(m = 2k) = \Pr(m = -2k) = 0.02$ for $k = 1, \ldots, 8$ indicating varying degrees of confounding among the subpopulation. "Shift confounding" represents a shift in the baseline hazard for all external controls. In particular, $x_3 = 1$. "Partial shift confounding" represents a shift in baseline hazard for the latter interval of survival time (i.e., $t > \tau_1$). In particular, $x_3 = I[t > \tau_1]$ is the indicator that t falls in the second of the two intervals which will be used in this illustration.

The values of β_3 used in the covariate effect $x_{3j}\beta_3$ range from $-\log(3)$ to $\log(3)$ representing hazard ratios between -3 and 3. The baseline hazard $h_0\left(t\middle|\boldsymbol{\lambda}\right)$ and the censoring distribution represented by the hazard $h_0(c|\boldsymbol{\lambda}^c)$ are taken as piecewise constant with $0 = \tau_0 < \tau_1 < \tau_2 = \infty$. To analyze properties of the case weights under specific intervals, the cutpoints used in the analysis were pre-specified. Similar operating characteristics were observed using our default suggestion of cutpoints induced by quantiles.

We consider an analysis which incorporates data from 200 subsampled RCT treated subjects, 100 subsampled RCT controls, and 100 subsampled external controls in an augmented analysis under different assumptions for the covariate x_3 and the magnitude of β_3. For the simulation studies, we find estimated values of $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^c$, γ, β_1, and β_2 from an analysis of the RCT subjects, using cutpoints $\boldsymbol{\tau}$ chosen to have an equal number of events in each interval.

We consider two scenarios for the prevalence of censoring in the external control population, both of which are based on modifications of the fitted value $\boldsymbol{\lambda}^c$ from the actual RCT data. "Low Censoring" will consider the baseline hazard for censoring to be $1.4 \cdot \boldsymbol{\lambda}^c$, and "High Censoring" will consider $0.9 \cdot \boldsymbol{\lambda}^c$. Note that the censoring distribution in the RCT and external controls are not assumed to be equivalent; the RCT data are used to provide an initial estimate $\boldsymbol{\lambda}^c$, which is further perturbed by the

multiplicative factors of {0.9, 1.4}. Since the number of external controls are fixed, the "Low Censoring" scenario is associated with more events and thus more information contained in the external control data, and the "High Censoring" scenario is associated with fewer events and less information.

The hypothesis under consideration is the one-sided hypothesis H_0: $\gamma \geq 0$ vs. H_1: $\gamma < 0$. This hypothesis was evaluated by computing the posterior probability of $\gamma < 0$ being less than 0.025. The true parameter values considered are $\gamma = 0$ (for the null hypothesis) and $\gamma = \log(0.73)$ (for the alternative hypothesis). The simulation study summarizes 10 000 repetitions per value of β_3 used to produce confounding. All simulations were performed using R version 4.1.2 (R Core Team, 2017). The case weighted power prior and the discounted case weighted power prior are compared to fixed weight power priors with weights $a_0 \in \{0, 0.5, 1\}$, which includes no borrowing ($a_0 = 0$) and full-borrowing ($a_0 = 1$), as well as the commensurate prior and the case weighted commensurate prior. Analyses using the commensurate prior are fit using the Hamiltonian Monte Carlo algorithm using STAN version 2.27.0 and `cmdstanr` version 0.4.0.

3.2 Distribution of case weights

Figure 1 shows the distribution of the case weights for the partial contamination and shift confounding scenarios averaged across simulated trials. For the untransformed weights, when $\beta_3 = 0$, the external controls are exchangeable with the RCT controls and the weights $a_{j,k}$ are uniformly distributed on the unit interval with a mean value of 0.5. As the magnitude of β_3 increases, the distribution of the case weights begins to differ from a uniform distribution.

The case weights transformed by the case weight shrinkage function $f_p(a_{j,k})$ are similar to the untransformed weights in their average, however there is less overall dispersion. Notice that while the magnitude of β_3 increases, the average case weight decreases, and the range of the 75th to the 90th percentile lengthens considerably. This is because the effect of the confounding on survival time increases as the magnitude of β_3 increases, causing lower case weights to be assigned to those observations most impacted. The transformed weights $f_p(a_{j,k})$ approach zero as $a_{j,k}$ approaches zero, so those observations most impacted by the confounding could be assigned an arbitrarily low case weight. It is this shrinking of the case weights around 0.5 that enables the case weighted power prior to have a controlled type I error rate; the transformed weights are closer to the fixed value of 0.5, which mimics a power prior with a fixed weight and a conservative type I error rate (see also Web Appendix D).

The case weights transformed by both the case weight shrinkage function and the uniform discounting function $f_p(a_{j,k})g_c(\overline{A})$ are similar to the transformed case weights $f_p(a_{j,k})$ for values of $|\beta_3|$ near 0, and are nearly equivalent when $\beta_3 = 0$. This is because there is little dataset-level incompatibility detected in the external controls. However, as β_3 increases, the average case weight drops substantially, as the difference between $\overline{A}$ from 0.5 increases. It is this drop in case weights that enables the discounted case weighted power prior to have a calibrated maximum type I error rate under shift confounding; for large levels of incompatibility in the external data, the amount of borrowing decreases substantially (see also Web Appendix D). Additional analysis of the case weights by amount of censoring and interval of time at risk is provided in Web Appendix F.

3.3 Operating characteristics

3.3.1 Partial contamination

Figure 2 shows the operating characteristics of type I error, power, and mean squared error for the case weighted power prior for the partial contamination scenario. The power at the null value of $\beta_3 = 0$ is the highest for full-borrowing, followed by the fixed weight power prior with $a_0 = 0.5$. However, all fixed weight power priors suffer precipitous drops in power as as $|\beta_3|$ increases, since incompatible external control information is incorporated to a fixed (i.e., static) degree. The case weighted power priors maintain a high level of power for all values of β_3 considered since external controls with observed incompatibility are dynamically down-weighted. All the fixed weight power priors explored are shown to have relatively lower MSE compared to the case weighted approaches for an interval of β_3 around 0, then sharply increase to levels higher than no borrowing as β_3 increases. Case weighting maintains a relatively low MSE for all values of β_3 considered. Among the dynamic borrowing approaches considered, the commensurate prior has the greatest reduction in power as $|\beta_3|$ increases, since the commensurate prior has no mechanism to dynamically weight individual outlying observations.

3.3.2 Shift confounding

Figure 3 shows the operating characteristics of the case weighted power prior for the shift confounding scenario. Case weighting maintains higher power than fixed weight power priors for $\beta_3 < 0$ by dynamically down-weighting the external controls with observation times that are observed to be incompatible with the RCT data. All fixed weight power priors are shown to have very high power for $\beta_3 > 0$, since this results in a downward bias in the estimated hazards for controls resulting in a upward bias in the estimated treatment effect. The case weighted power prior has lower power in this case due to the down-weighting of the incompatible external controls, demonstrating that the case weighted power prior does not uniformly increase power in all scenarios. As in the case of partial contamination, all fixed weight power priors have relatively low MSE for an interval of β_3 around zero, while the case weighted power prior has relatively low MSE for all values of β_3 considered.

The commensurate prior and case weighted commensurate prior behave similarly in terms of the type I error rate and power, while the commensurate prior has relatively lower MSE as $|\beta_3|$ increases. The commensurate priors with the chosen specification of hyperprior on the variance for the drift parameter have less spread around no borrowing for type I error and power.

3.3.3 Partial shift confounding

The operating characteristics from the partial shift confounding scenario are generally similar to those of the shift confounding scenario displayed in Figure 3 (see Web Appendix F). The magnitude of the type I error rate and power differences from the no

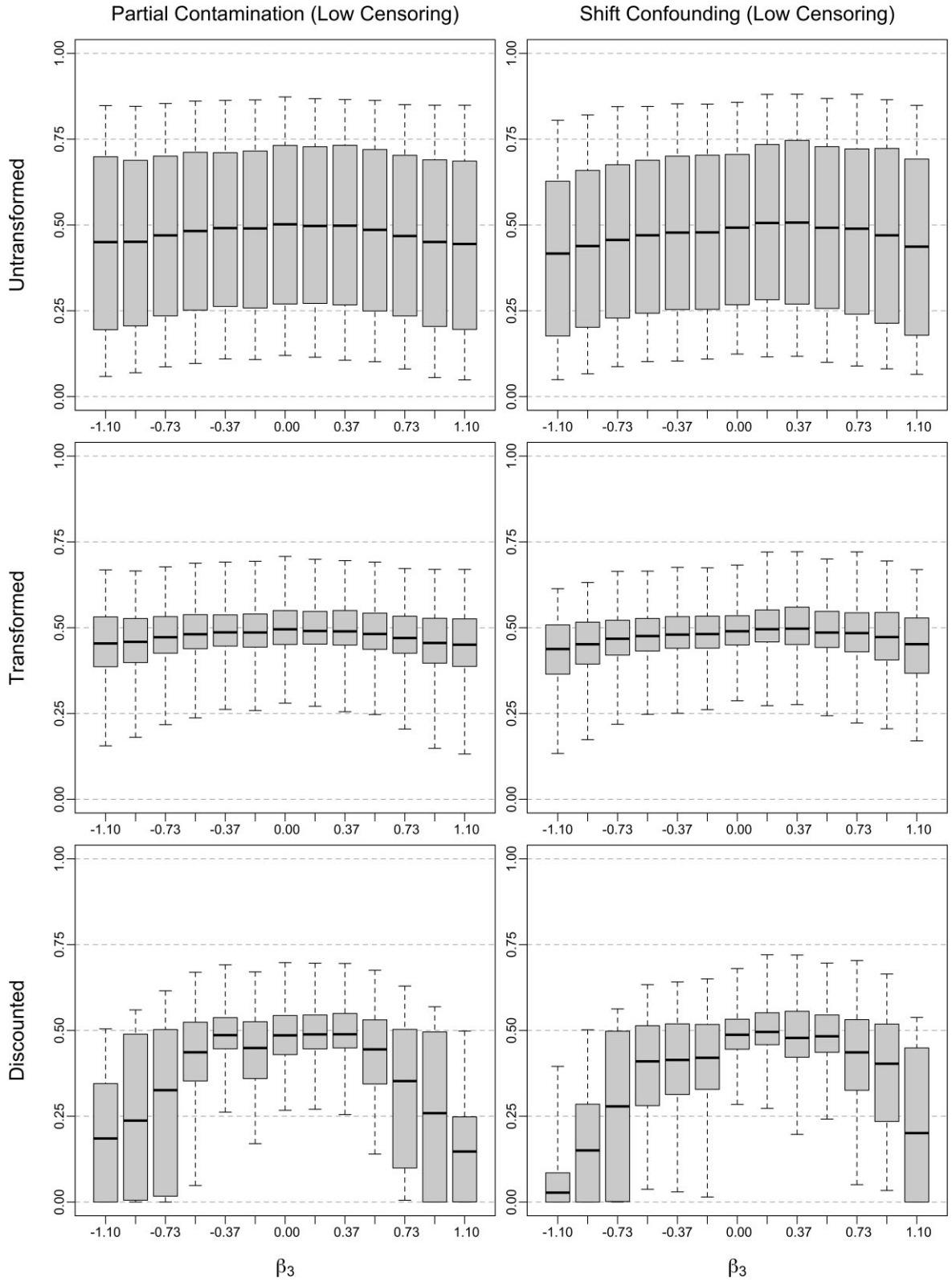

FIGURE 1 Case weight distributions by data generation scenario and analysis method. Untransformed = untransformed case weighted power prior, transformed = case weighted power prior; discounted = discounted case weighted power prior with maximum type I error rate under shift confounding calibrated at 0.15. Boxplot with mean, inter-quartile range, and 10th/90th percentiles of case weights.

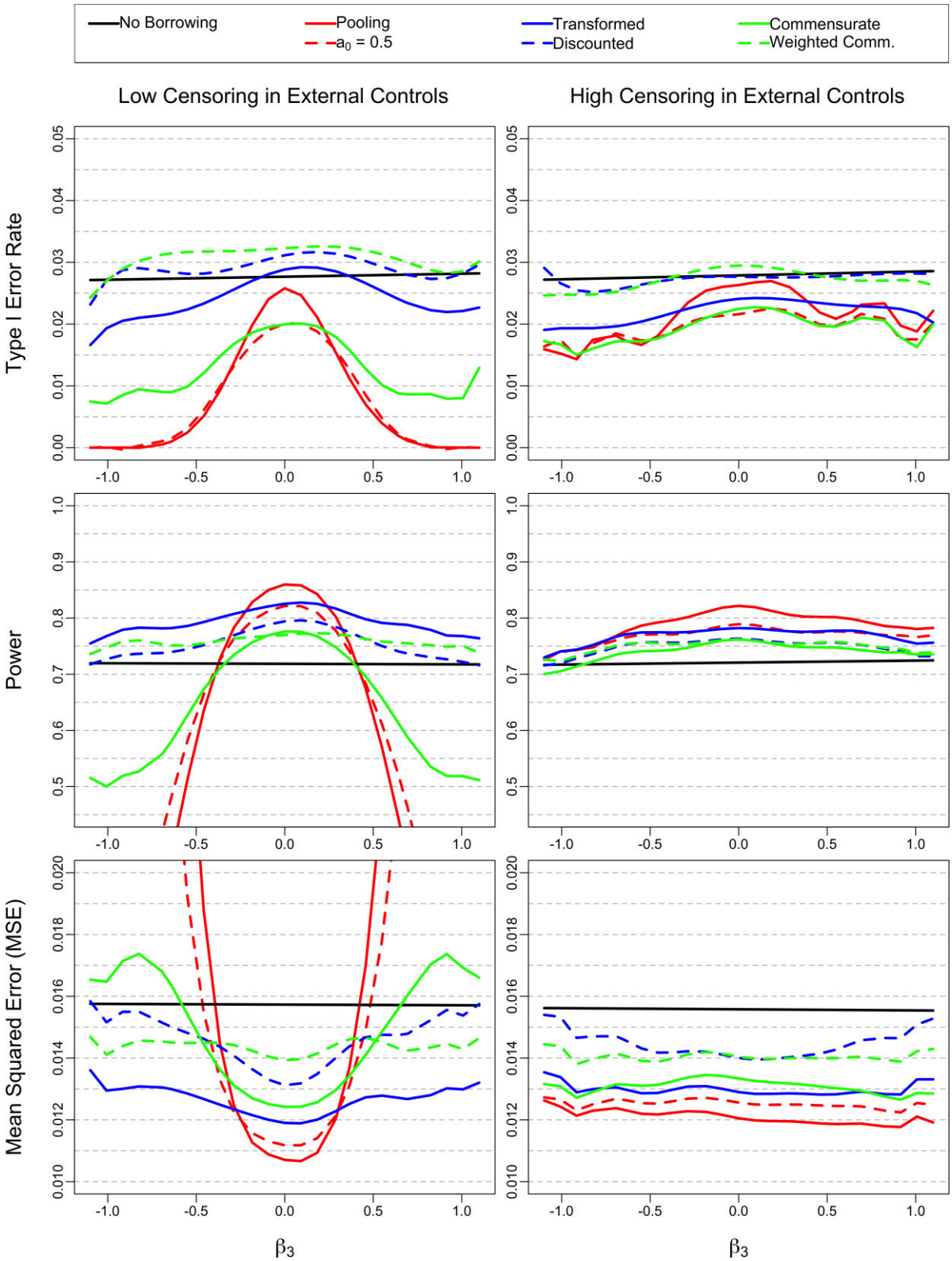

FIGURE 2 Partial contamination scenario; selected operating characteristics by analysis method. Transformed = case weighted power prior; discounted = discounted case weighted power prior with maximum type I error rate under shift confounding calibrated at 0.15.

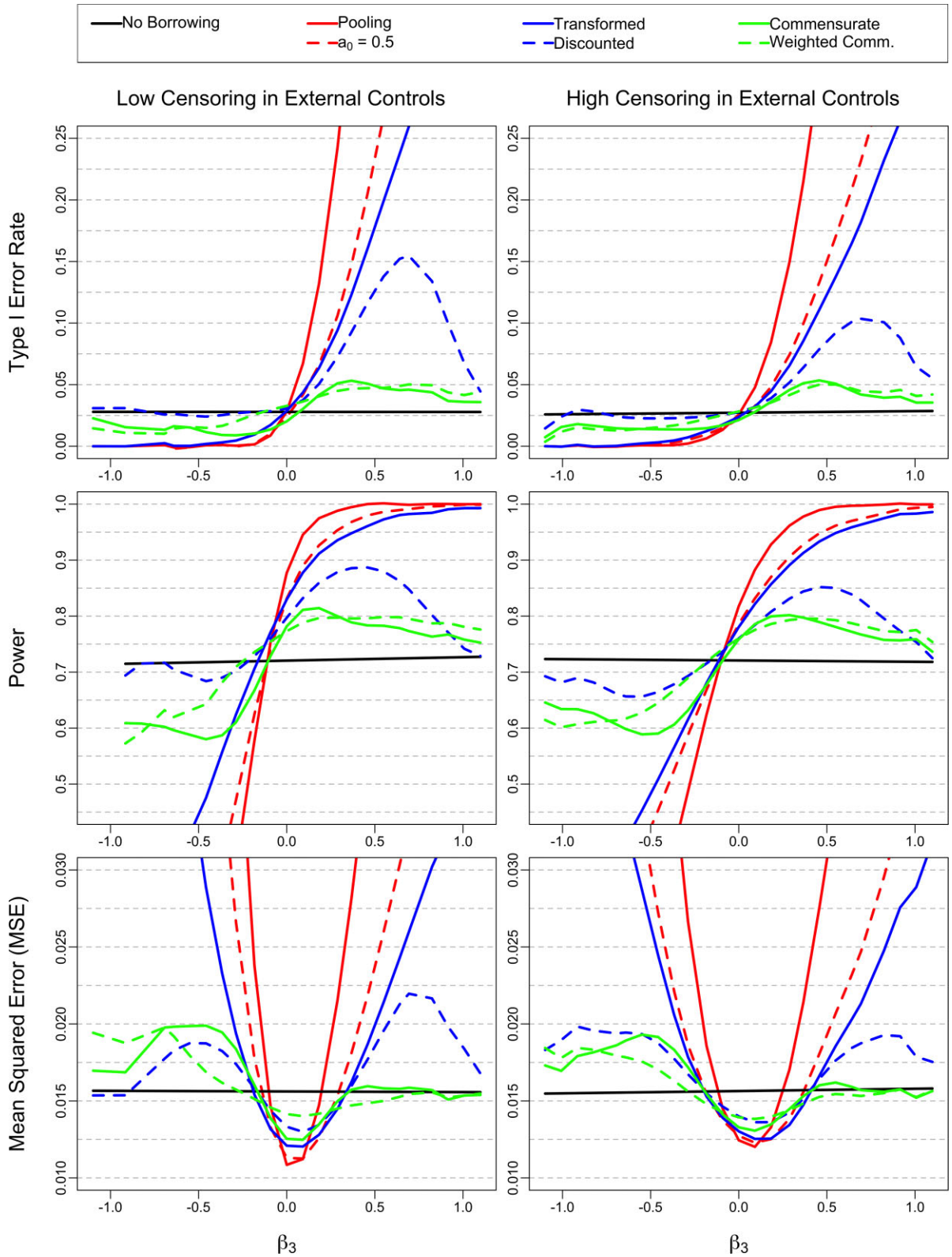

FIGURE 3 Shift confounding scenario; selected operating characteristics by analysis method. Transformed = case weighted power prior; discounted = discounted case weighted power prior with maximum type I error rate under shift confounding calibrated at 0.15.

borrowing case are shown to be less than those from the shift confounding scenario for all the power priors. This is to be expected since the partial shift confounding scenario corresponds to a lesser degree of incompatibility than the shift confounding scenario.

3.4 Choosing number of segments for baseline hazard

It is necessary to give thoughtful consideration to the number of segments used for the baseline hazard in the analysis model. Web Table 5 shows that the lowest BIC occurs when the number of segments for the baseline hazards in the generating model matches the number of segments used in the analysis model. Consequently, these situations produce the more accurate results for the treatment effect estimation. For example, when $\beta_3 = 0$ (i.e., compatible external controls) and $K_G = 3$ segments are used in the generating hazard, the closest estimated hazard ratio to the generating value of 0.73 is observed when $K_M = 3$ segments are used in the analysis model.

3.4.1 Vary proportion of external controls

We explore the impact of modifying the number of external controls from 100 to either 200 or 50 while keeping the same amount of 200 subjects randomized to treatment and 100 randomized controls. Modifying the number of external controls fundamentally alters the study's operating characteristics (see Web Appendix F). For example, the maximum power in the homogeneous case with 100 external controls is 0.86, which increases to 0.92 with 200 external controls and decreases to 0.82 with 50 external controls. Relatedly, the type I error is much more sensitive to confounding with more external controls, and less sensitive to confounding with fewer external controls. These are unavoidable consequences of having varying amounts of information informing the control group rather than distinctive shortcomings of the proposed methods.

Although the operating characteristics of the designs are altered, the proposed methods maintain their comparative advantages to the benchmark methods of no borrowing and pooling. When the number of external controls is modified, the type I error rate is still preserved. For the partial contamination scenario, the proposed methods still maintain higher power and lower MSE across the magnitude of the confounding variable β_3. For the shift confounding scenario, the proposed methods are more robust to confounding than the pooling method, and provide increases in power and decreases in MSE relative to no borrowing when the unobserved confounding is minimal.

3.4.2 Study impact of model misspecification

To study the impact of model misspecification, we consider delayed separation situations where the treatment effect is only present after a delay of 50 or 100 days. These modifications alter the operating characteristics of the study: for example, maximum power is decreased from 0.86 to 0.82 with a 50-day delay and 0.77 with a 100-day delay (see Web Appendix F). Still, the proposed methods maintain their comparative advantages to the benchmark methods of no borrowing and pooling, such as higher power and lower MSE in the partial contamination scenario, and better robustness to shift confounding than pooling with increases in power and decreases in MSE relative to no borrowing when the unobserved confounding is minimal.

4 REAL DATA EXAMPLE

We consider testing the adaptive borrowing method using all subjects from the real datasets, which served as motivation for the simulation studies (i.e., 850 randomized subjects and 526 external controls). Figure 4a shows observed differences across the RCT and external datasets. The Kaplan-Meier curves show that the treatment arm has improved survival relative to the randomized control arm, which in turn has improved survival relative to the external control arm. Figure 4b shows fitted model coefficients for the RCT and external data, which implies compatibility in that the covariate effects of age and sex are highly similar between the datasets, and also implies incompatibility in that the baseline hazard components are lower for the RCT indicating improved survival (although the estimated cofficients have overlapping 95% confidence intervals). An examination of the compatibility weights for the external controls in Figure 4c demonstrates limited deviation from the anticipated uniform distribution. Table 1 shows the estimate for the treatment effect and model fit diagnostics using the case weighted power prior. It is shown that when $\boldsymbol{\tau}$ has 3 cutpoints, the BIC is the lowest, and the estimated hazard ratio associated with the treatment effect is 0.650. This estimated hazard ratio is equivalent to the estimated hazard ratio for the pooling method, although the pooling method has a slightly narrower credible interval.

5 DISCUSSION

The case weighted power prior provides a novel strategy for the incorporation of RWD into an analysis of RCT data. The case weights provide a framework for comparatively more robust estimation of effects of interest compared to fixed weight power priors in scenarios where there is systematic incompatibility between RCT controls and external controls due to unmeasured confounding. Using predictive distributions (e.g., Box's p-value) provides an intuitive metric of compatibility for comparing external control data to RCT data. This method increases power in the case of compatibility but also succeeds in reducing the influence of incompatible external controls and limiting the increase in the type I error rate. Since each external control subject is assigned their own compatibility weights, this method can be directly applied to incorporating different sources of external controls into a single hybrid analysis. Addressing compatibility for time-to-event outcomes is an involved process, and we are aware of no other methods which can account for a partial shift in baseline hazard among external controls. A straightforward application of this methodology would be toward non-survival outcomes (e.g., normally distributed study endpoints).

A common model for the observation times is assumed for both the RCT and external control data. This approach is binding, but essential. If we are to evaluate how well RCT data predicts the external control data, we must have a common model

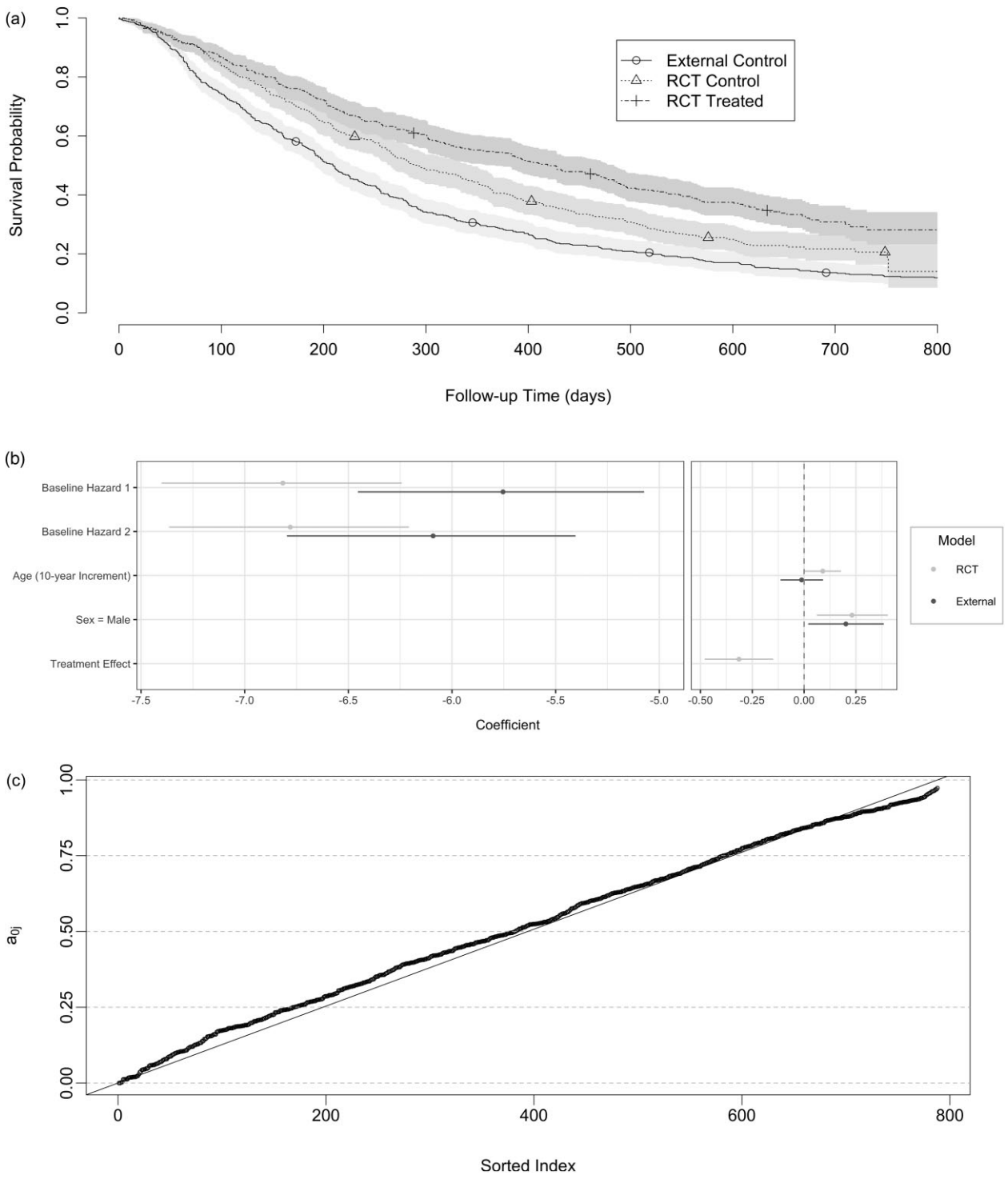

FIGURE 4 (a) Kaplan-Meier curves for RCT and external data. (b) Fitted model coefficients for RCT and external data. (c) Compatibility weights for NSCLC RWD.

to translate between them. From this perspective, the pretext for borrowing from external controls is established at the outset using criteria such as that in Pocock (1976), and deviations from a shared proportional hazards model found from diagnostics such as scaled Schoenfeld residuals would not be used to justify or nullify the basis for borrowing. These possible deviations from a shared model are implicitly incorporated into the case weights using our compatibility assessments.

The capacity to detect prior-data conflict among external controls is heavily dependent on the total amount of information in the RCT, which in this context is determined by the number of events. This needs to be considered for this

TABLE 1 Treatment effect estimation and model fit diagnostics for OAK and NSCLC RWD datasets. K_M: number of baseline hazard segments used in the analysis model, $\overline{a_0}$: average case weight, HR: hazard ratio, CI W.: credible interval width, BIC: Bayesian Information Criterion.

		Adaptive			No borrowing			Pooling		
K_M	$\overline{a_0}$	HR	CI W.	BIC	HR	CI W.	BIC	HR	CI W.	BIC
1	0.474	0.670	0.197	14922.8	0.732	0.242	14931.7	0.653	0.182	14921.8
2	0.536	0.662	0.194	14940.1	0.730	0.242	14933.0	0.657	0.183	14956.6
3	0.544	0.650	0.190	14916.1	0.728	0.241	14935.7	0.650	0.181	14911.9
4	0.553	0.652	0.190	14975.2	0.729	0.242	14982.5	0.649	0.181	14992.7
5	0.556	0.651	0.190	14918.4	0.727	0.241	14942.7	0.643	0.180	14913.3

methodology to be applied to study designs with interim analyses where limited numbers of events are available in the RCT. Specifically, there needs to be an adequate number of events available to characterize the predictive distribution which provides the basis for the compatibility assessment. In the limit (i.e., no events in the RCT) the external controls are evaluated against the prior predictive distribution for the RCT data which does not reflect data-driven estimates for parameters related to survival. Thus, for this method to be useful, there must be sufficient numbers of events in the RCT for the likelihood to dominate the analysis prior. Formalizing this recommendation for trial design contexts is an area of future research.

Future work could involve testing robustness of the case weighted power priors under additional types of confounding between the RWD and RCT data, such as combinations of partial contamination and shifts in baseline hazards. Data quality issues relating to RWD remain a persistent challenge, including difficulty in defining time zero for an external control, which could result in immortal time bias favoring the RCT group (Burcu et al. 2020).

While there is interest in obtaining drug approvals using single-arm studies, our method determines case weights for the external controls by assessing how well the predictive distribution (based on the RCT) for their observation time data aligns with the actual observed data for the external controls, which is the uniqueness of hybrid control trials and analyses. In order for the proposed approach to be feasible, one must have some controls in the prospective trial. Thus, as constructed, this method would not be directly applicable to purely externally controlled trials.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Appendices, Tables, and Figures referenced in Sections 2 and 3, as well as code, are available with this paper at the Biometrics website on Oxford Academic.

FUNDING

This work was partially supported by U.S. Food and Drug Administration grant HHS U01 FD007206 (Lieberman, Pang, Li, Zhu). E.K. was partially supported by the National Cancer Institute T32 training grant [5T32CA096520-15].

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The data used in Section 3 are included in the GitHub repository (https://github.com/kwiatkowski-evan/case-weighted-power-priors), which includes the programs and other resources needed to reproduce the analyses in this paper. The data used in Section 4 cannot be shared publicly for the privacy of the individuals that participated in the study.

REFERENCES

Birnbaum, B., Nussbaum, N., Seidl-Rathkopf, K., Agrawal, M., Estevez, M., Estola, E. et al. (2020). Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. *arXiv*, 2001.09765. Accessed 6 Jul. 2022.

Bouaziz, O. and Nuel, G. (2016). L0 regularization for the estimation of piecewise constant hazard rates in survival analysis. *Applied Mathematics*, 8, 377–394.

Box, G. E. P. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, 143, 383–430.

Burcu, M., Dreyer, N. A., Franklin, J. M., Blum, M. D., Critchlow, C. W., Perfetto, E. M. et al. (2020). Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms. *Pharmacoepidemiology and Drug Safety*, 29, 1228–1235.

Chen, W. C., Lu, N., Wang, C., Li, H., Song, C., Tiwari, R. et al. (2022). Propensity score-integrated approach to survival analysis: leveraging external evidence in single-arm studies. *Journal of Biopharmaceutical Statistics*, 32, 400–413.

Dejardin, D., Delmar, P., Warne, C., Patel, K., van Rosmalen, J. and Lesaffre, E. (2018). Use of a historical control group in a noninferiority trial assessing a new antibacterial treatment: A case study and discussion of practical implementation aspects. *Pharmaceutical Statistics*, 17, 169–181.

Evans, M. and Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, 1, 893–914.

Fu, C., Pang, H., Zhou, S. and Zhu, J. (2023). Covariate handling approaches in combination with dynamic borrowing for hybrid control studies. *Pharmaceutical Statistics*, 22, 619–632.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*, Taylor and Francis, New York.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J. and Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67, 1047–1056.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, 15, 46–60.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis*, Springer, New York.

Kwiatkowski, E., Andraca-Carrera, E., Soukup, M. and Psioda, M. A. (2022). A structured framework for adaptively incorporating external

evidence in sequentially monitored clinical trials. *Journal of Biopharmaceutical Statistics*, 32, 474–495.
Lek, K. and Van De Schoot, R. (2019). How the choice of distance measure influences the detection of prior-data conflict. *Entropy*, 21, 1–17.
Lewis, C. J., Sarkar, S., Zhu, J. and Carlin, B. P. (2019). Borrowing from historical control data in cancer drug development: A cautionary tale and practical guidelines. *Statistics in Biopharmaceutical Research*, 11, 67–78.
Ma, X., Long, L., Moon, S., Adamson, B. J. S. and Baxi, S. S. (2020). Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR. *medRxiv*, 2020.03.16.20037143. Accessed 6 Jul. 2022.
Neuenschwander, B., Capkun-Niggli, G., Branson, M. and Spiegelhalter, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7, 5–18.
Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29, 175–188.
Psioda, M. A. and Ibrahim, J. G. (2018). Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics*, 20, 400–415.
Psioda, M. A., Soukup, M. and Ibrahim, J. G. (2018). A practical Bayesian adaptive design incorporating data from historical controls. *Statistics in Medicine*, 37, 4054–4070.
Psioda, M. A. and Xue, X. (2020). A Bayesian adaptive two-stage design for pediatric clinical trials. *Journal of Biopharmaceutical Statistics*, 30, 1091–1108.
R Core Team (2017). A language and environment for statistical computing. *R Foundation for Statistical Computing*, https://www.R-project.org/.
Rittmeyer, A., Barlesi, F., Waterkamp, D., Park, K., Ciardiello, F., von Pawel, J. et al. (2017). Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): A phase 3, open-label, multicentre randomised controlled trial. *Lancet (London, England)*, 389, 255–265.
Schmidli, H., Häring, D. A., Thomas, M., Cassidy, A., Weber, S. and Bretz, F. (2019). Beyond randomized clinical trials: Use of external controls. *Clinical Pharmacology and Therapeutics*, 107, 806–816.
Ventz, S., Lai, A., Cloughesy, T. F., Wen, P. Y., Trippa, L. and Alexander, B. M. (2019). Design and evaluation of an external control arm using prior clinical trials and real-world data. *Clinical Cancer Research*, 25, 4993–5001.
Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N. et al. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13, 41–54.
Wang, C., Li, H., Chen, W. C., Lu, N., Tiwari, R., Xu, Y. et al. (2019). Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *Journal of Biopharmaceutical Statistics*, 29, 731–748.
Wang, X., Suttner, L., Jemielita, T. and Li, X. (2022). Propensity score-integrated Bayesian prior approaches for augmented control designs: A simulation study. *Journal of Biopharmaceutical Statistics*, 32, 170–190.

Received: July 6, 2022; Revised: December 31, 2023; Accepted: February 27, 2024